DOCUMENT RESUME

ED 057 087 TM 000 930

AUTHOR Cooley, William W.
TITLE Methods of Evaluating School Innovations.
PUB DATE 3 Sep 71
NOTE 30p.; Paper presented at the 79th Annual Convention of the American Psychological Association, Washington, D.C., September 1971

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Academic Achievement; *Classroom Research; Criterion Referenced Tests; Curriculum Development; Early Childhood Education; Educational Innovation; Educational Programs; *Evaluation Methods; Experimental Curriculum; *Formative Evaluation; Individual Differences; *Individualized Instruction; Input Output Analysis; *Instructional Innovation; Instructional Materials; Program Evaluation
IDENTIFIERS *Individually Prescribed Instruction; IPI; Learning Research and Development Center; LRDC; Wide Range Achievement Tests; WRAT

ABSTRACT

This evaluative research is concerned with specific educational programs which attemot to adapt instruction to individual differences. Attention is limited to the Frick School, a large urban Pittsburgh school in which the Learning Research and Development Center develops its new educational programs, and to the Follow-Through network where these programs are used by the center to study processes involved in dissemination of new educational innovations. (CK)

9/3/71

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

ED057087

Methods of Evaluating School Innovations[1]

William W. Cooley

This past year I was asked to prepare a book of readings for the new AERA series, which was to illustrate the current state of evaluation in curriculum development. As some of you probably would guess, my search for actual published examples of evaluation was disappointing. The field does not have a significant literature. There is an abundance of papers about evaluation models and strategies, and "how to-do-it recipes." There are many attempts at developing a taxonomy of evaluation-type activities. But there is a great scarcity of publicly available publications which report the procedures and results of actual evaluation studies. The results either never reach the stage of printed word, or if they do, they tend to be in the form of large "telephone books" which are of limited distribution and primarily end up as part of the wall structure of the U.S. Office of Education.

It does not seem probable that evaluation studies will develop and improve if reports are not made generally available and critically examined by other scholars in the field. Thus, as I began to prepare this paper, I concluded that what the world needs is not another paper about evaluation but rather a description of an evaluation, a paper which illustrates a

---

[1]Invited Address to 79th Annual Convention, American Psychological Association, Washington, D.C., September 3, 1971.

TM 000 630

researcher struggling with data in an attempt to provide unambiguous information to others regarding the worthwhileness of new educational materials and procedures.

The only thing I want to say <u>about</u> evaluation is that evaluating school innovations should be nothing more or less than good research, in the sense that research is the process whereby one attempts to provide evidence regarding the validity of a proposition. It is my conviction that evaluative research differs from basic and applied research only in the nature of the propositions under study and in how they are initially formulated. In basic research the propositions to be studied come from theory and a network of other related propositions. In applied research the propositions to be investigated come more out of the application of science, and are formulated when the validated principles which that science has produced to date are inadequate to a particular application of that science. Evaluative research, as one type of applied research, attempts to assess the validity of propositions regarding <u>particular</u> programs and procedures, rather than propositions about variables which may be common to many programs.

The context of my illustrations is the Learning Research and Development Center (LRDC) at the University of Pittsburgh and the specific instructional materials and procedures under development at that Center. Thus, this evaluative research is concerned with specific educational programs which attempt to adapt instruction to individual differences. The objective of the research is to provide information regarding

the validity of propositions about LRDC's educational programs. The propositions, together with data regarding their validity, are designed to provide information to others regarding the worthwhileness of the new programs, and to provide information to developers regarding the relative strengths and weaknesses of components of the programs.

There are four settings in which LRDC is able to study the fruits of its development efforts. Perhaps the most widely known of these settings is the Oakleaf School, a small elementary school in a suburb of Pittsburgh, where Individually Prescribed Instruction was introduced seven years ago (Lindvall and Bolvin, 1967). A second setting is the network of field test schools established by Research for Better Schools, Inc. (RBS), a regional laboratory in Philadelphia. RBS has been engaged in the dissemination of LRDC products developed at the Oakleaf School since 1966. A third setting is the Frick School, a large urban school in the City of Pittsburgh in which the Center has been developing programs for the past four years. Programs that have been developed and tested in the Frick School are then moved out into a fourth setting known as our Follow-Through network. This past year, four school systems, with Follow-Through funding, selected and implemented the programs we developed in the Frick School in their own elementary schools. The Center is engaged in this Follow-Through network so that it can study the processes involved in dissemination of new educational innovations. Lindvall and Cox (1970) and RBS (1971) provide and summarize reports of evaluations done of the

Oakleaf-RBS enterprises. I shall restrict my concern today to the Frick and Follow-Through settings.

In the Frick School, LRDC has been developing an individualized program which involves the development of an instructional plan for each child based upon the results of individually administered criterion-referenced tests, a prescription system which implements that individual plan on a day-to-day basis, a redefined classroom teacher's role which emphasizes testing, tutoring, and traveling, resulting in a structured curriculum in basic perceptual, reading, and arithmetic skills. This then is complemented by an exploratory curriculum, in which the child freely selects open-ended learning activities in such areas as creative and language arts, socio-dramatic play, science, and social studies.

The Frick program began in preschool and kindergarten in 1968-69, added first grade in 1969-70, second grade this past year, and will include preschool through third grade for the coming school year. The Follow-Through network began with three school systems in 1969-70, added a fourth school system this past year, and goes to seven this Fall. We plan to limit this network to the present seven systems, which is large enough to study the dissemination problem and evaluate our products, yet small enough to be manageable as a research and development enterprise.

Most evaluations that have actually included data on students have been efforts at testing the validity of the proposition that the innovation

under investigation is better than some alternative approach, what American TV calls "brand X." The investigator then proceeded to define "better" by some kind of standardized achievement measure or set of measures. He defined brand X by establishing some control schools or classrooms, and then compared the resulting means. If there were no differences, the innovators were convinced that the results were not valid and continued to try to show how their innovation was superior to other programs available in the schools. If the results of the comparison indicated that the innovation was superior to brand X, the innovators congratulated themselves and the evaluator on a job well done. Those who were skeptical of the innovation found flaws in the design and proceeded to debunk the validity of the proposition.

Let me turn to some Frick School results to illustrate this point. In order to contrast LRDC's program to the school's previous program, control groups were established in Frick by taking advantage of the fact that we were adding one grade each year to our program. Table 1 illustrates the general design worked out by Wang, Resnick, and Schuetz (1970). Thus, we began the experimental program in preschool and kindergarten, and to establish control groups, tested two grades ahead of the program as it moved up from grade to grade. No significant differences in achievement were observed between controls of a given grade from year to year. Also, no differences were found on variables known to be related to

Table 1

Experimental (E) and Control (C) Groups for Frick School

Year	Grade						
	Pre-school	Kinder-garden	First	Second	Third	Fourth	Fifth
1968-69	E	E	C	C	--	--	--
1969-70	E	E	E**	C*	C	--	--
1970-71	E	E	E**	E*	C	C	--
1971-72	E	E	E	E	E	C	C

*Contrast illustrated in Table 2.
**Contrast illustrated in Table 3.

achievement but which could not have been affected by the program, such as family socio-economic status. Therefore, it is reasonable to assume that from year to year, children at a given grade level were random samples from a common population.

Table 2 results show that the new program made statistically significant increases in all three achievement areas measured by the Wide Range Achievement Test (WRAT) (Jastak, Bijou, & Jastak, 1965) for the second grade. The spelling results were of special interest to our reading developers because we do not try to teach spelling directly, but expect it as a by-product of how reading is taught.

The norming information on the test allows us to get some idea of how much of a change this represents in terms of grade equivalents. The resulting differences indicate an increase of seven months in reading and four months in spelling and arithmetic.

The results in Table 3 show the effect of the changes made between the first and second approximations to our first-grade program. Evaluation of the program in the Frick first grade during 1969-70 led to the changes which were implemented in the Fall of 1970. The first and second year contrast provides information useful in monitoring the development effort. Changes in the program cannot be claimed as improvements until the effects of these changes are known. The significant improvements here provide encouragement to the developers in that they seem to be on the right

Table 2

Second Grade Comparisons Before and After LRDC Program
(Wide Range Achievement Test)

	"Before" (Spring 1970) (N = 98)	"After" (Spring 1971) (N = 116)
Reading		
Mean (raw)	41.45	49.91
St. dev. (raw)	9.69	13.80
Grade equiv.	2.2	2.9
	F = 25.96; ndf = 1 and 212; prob. < .001	
Spelling		
Mean (raw)	26.20	28.72
St. dev. (raw)	5.08	5.44
Grade equiv.	1.9	2.3
	F = 8.51; ndf = 1 and 212; prob. < .01	
Arithmetic		
Mean (raw)	23.40	25.22
St. dev. (raw)	2.85	3.42
Grade equiv.	2.2	2.6
	F = 17.62; ndf = 1 and 212; prob. < .001	

Table 3

First Grade Comparisons Following Changes in the LRDC Program
(Wide Range Achievement Test)

	After 1st Year (Spring 1970) (N = 143)	After 2nd Year (Spring 1971) (N = 124)
Reading		
Mean (raw)	34.27	41.37
St. dev. (raw)	10.32	11.85
Grade equiv.	1.7	2.2
	F = 27.41; ndf = 1 and 265; prob. < .001	
Spelling		
Mean (raw)	20.64	25.53
St. dev. (raw)	4.65	5.77
Grade equiv.	1.3	1.7
	F = 58.89; ndf = 1 and 265; prob. < .001	
Arithmetic		
Mean (raw)	22.36	23.98
St. dev. (raw)	3.24	2.58
Grad. equiv.	2.1	2.4
	F = 20.03; ndf = 1 and 265; prob. < .001	

track. In addition to gains from one experimental year to the next, the first-graders are now performing as well as the second-graders were before our program began (compare the means of the second column of Table 3 with the means of the first column of Table 2).

But are these results useful to others who wish to judge the worthwhileness of our new program? Certainly they are encouraging to the developer. Innovations do not always yield increases in means, even though one seldom finds such negative results in the literature. Do these results convince you that this program belongs in the elementary schools of your community? Certainly not! Yet insofar as student achievement data have been included in actual school evaluations at all, similar results have been offered as proof of a program's worthwhileness.

Many inadequacies of such results immediately come to mind:

1. Restricting results to one experimental school does not show how a program would work in the field in a variety of school settings.

2. Restricting contrasts to one achievement test, known to the developer, tends to produce a credibility gap in the mind of the skeptical consumer.

3. Statistical evidence alone never convinces anyone of anything. The burden is on the innovator to show how the new program works, as well as how well it works. In the absence of a convincing mechanism, statistical results tend <u>not</u> to be believed. The cigarette controversy is

a classical example of this point. The statistical trend relating cigarette smoking and cancer was known for a long time, but few took the results seriously until it was also possible to show how cigarette smoking can produce cancer. The initial trend was important, however, because it stimulated and guided the search for how.

In order to overcome the shortcomings which result from restricting evaluation studies to one experimental school, we can turn to our Follow-Through network of schools. But as the new programs are moved out from the experimental school "hot-house" into the field, new problems arise. How can we be sure the model we built is the one operating in the classroom? After a teacher is trained in the new procedures, and the new materials are delivered to the classroom, she closes the door and does her thing, and her thing may not be our thing. What is needed is a method of determining the degree to which each classroom actually implements the instructional model, and a method of relating degree-of-implementation data to the achievement results of children in each classroom. By turning to the classroom as the unit of analysis, it may be possible to solve this problem and contribute to an understanding of the essential characteristics of the instructional model.

Most previous evaluation studies of new curricula or new instructional models have used analysis of variance (ANOVA) as the primary statistical tool. More recent efforts have extended this to the multivariate generalization

(MANOVA), but the general design is still the same--two or more grossly defined educational treatments are contrasted on one or more achievement criteria, with experimental or statistical control of initial student differences. Neither the developer nor the potential consumer has learned much from such studies.

Since a convincing evaluation study must include a variety of classroom settings, and since these classrooms will vary in the degree to which different aspects of the instructional model are implemented, establishing dimensions for measuring degree of implementation and using the classroom as the unit of analysis in a correlational model seem to be promising approaches to explore.

Three sets of variables need to be included: (1) student entering behavior (input), (2) dimensions of instructional treatments (process), and (3) end-of-year student achievement (output). Of course, the main reason for using the classroom as the unit of analysis is because the process measures are characteristic of classrooms. However, another important feature of this approach is that it allows one to consider the effects of differences in classroom input distributions on output, and the ways in which treatment variation may be related to output distributions. This is accomplished by reducing each measure of student input or output to four statistics for each classroom: mean (M), standard deviation (s), skewness (g_1), and kurtosis (g_2). Figure 1 shows a frequency polygon

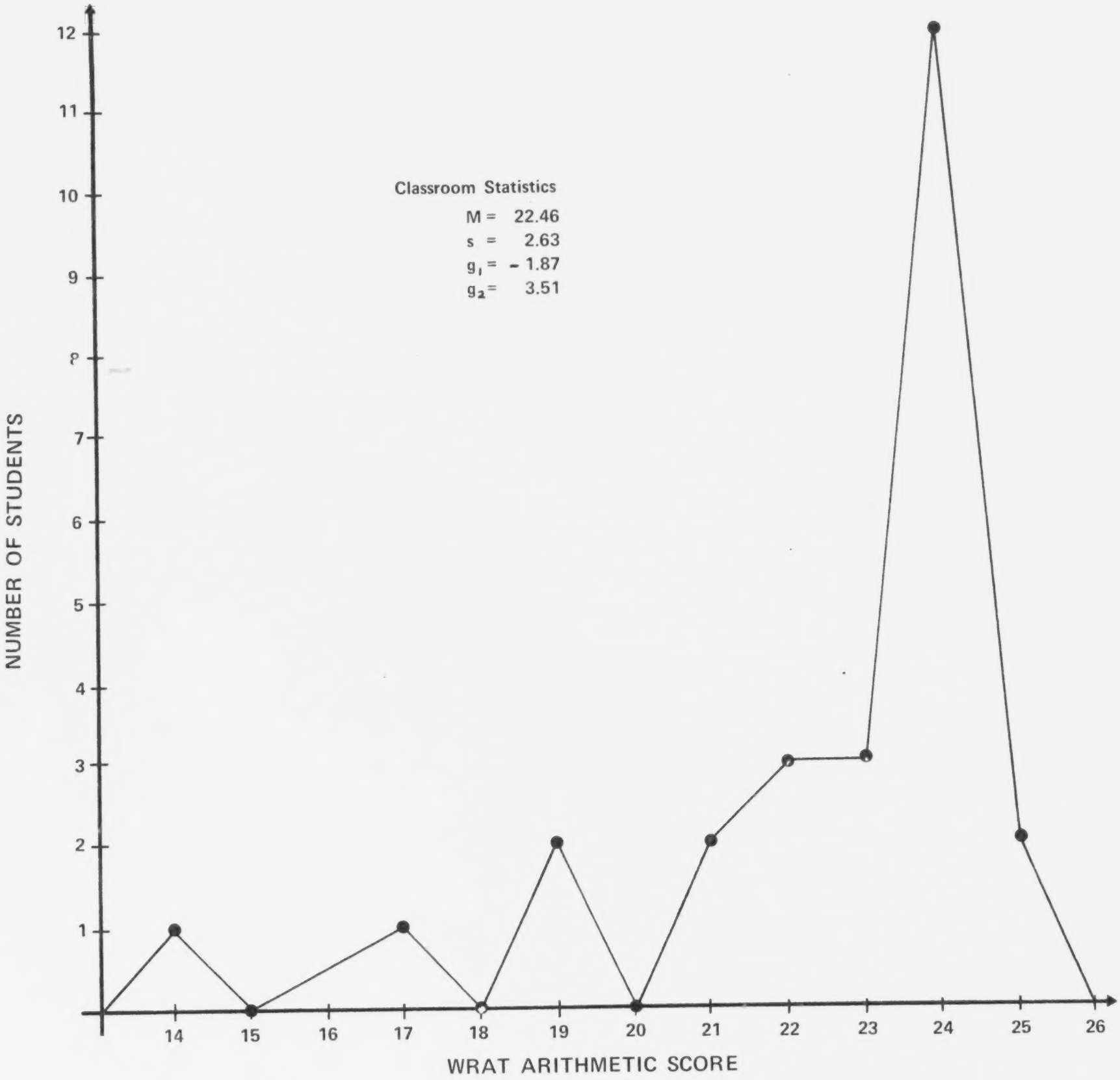

Figure 1

WRAT Arithmetic Distribution for Classroom 1114
(N = 26)

and the four statistics for one Frick classroom. The information regarding the negative skewness and peakedness of the distribution, as well as its location and general dispersion, is clearly preserved in these four values. Wiley (in Wittrock & Wiley, 1970) has suggested the usefulness of this approach, and Lohnes (1971) provides a good illustration in his reanalysis of the Cooperative Reading Study data. Let me try to make all of this clearer by illustrating this with some small examples using actual data from Frick and Follow-Through classrooms.

One dimension of student input is the placement test in our quantification curriculum (see Resnick, Wang, & Kaplan, 1970). A similar measure of student output is the arithmetic score on the WRAT. The scores on these two measures for 1,500 students can be converted to eight measures on 57 classrooms, the four classroom statistics based on quantification curriculum placement as input measures, and the four WRAT statistics as output measures.

Before bringing in the classroom treatment measures, it would be useful to examine the relationships among these eight input and output measures. Rather than stare at a correlation matrix of 64 elements, canonical correlation provides a good summary of how the input measures are related to the output. Table 4 summarizes the results of a canonical correlation analysis between the four input measures and the four output measures.

Table 4

Canonical Correlations Between Fall and Spring Measures
(N = 57 Classrooms)

Classroom Statistics	Mean	St. Dev.	Canonical Structure	Canonical Coefficients	
			INPUT		
Fall Quantification					
Mean	7.12	8.12	.82	.92	Variance Extracted = .37
St. dev.	5.90	5.84	.53	-.29	
Skewness	1.11	1.11	-.66	-.85	Redundancy = .20
Kurtosis	1.84	3.75	-.25	.66	
			OUTPUT		
Spring WRAT Arithmetic					
Mean	19.92	3.28	.99	.93	Variance Extracted = .33
St. dev.	3.17	1.01	-.57	-.12	
Skewness	-.49	.61	-.11	-.22	Redundancy = .18
Kurtosis	.59	1.48	.09	-.16	

Canonical Correlation = .73

Chi Square = 50.12

ndf = 16

$p < .001$

Other possible canonical relationships not significant at .05 level.

Only one of the possible four canonical relationships was significant at the .05 level. The canonical structure and the coefficients for that largest relationship show that a factor loaded positively on means and standard deviations and loaded negatively on skewness at input time correlates .73 with a factor primarily defined by the means at output time. Thus, the shape as well as the mean of the Fall distribution of students seems to affect the classes' Spring mean achievement, but the shape of the Spring achievement distributions is relatively unrelated to Fall input measures. That is, the amount of dispersion, skewness, and kurtosis in the Spring is related to Fall measures only insofar as it is related to the Spring means. Therefore, something other than input differences seems to be determining the shape of the Spring distributions.

The first canonical factor extracts about one-third of the variance from each of the two sets of variables (.37 and .33). The variance extracted together with the canonical correlation allows us to estimate the redundancy of output given input. A redundancy coefficient of .18 indicates that 82 percent of the total output variance is not explained by that first input factor.[2] This leaves a lot of output variance to be explained by something other than input variance.

Although canonicals between input and output may themselves be interesting, it is necessary to introduce a third set of measures, the process

---

[2] See Cooley and Lohnes (1971) for a discussion of this redundancy coefficient.

dimensions, into the analysis. First, let me describe the general nature of these measures of process, which we also call treatment or degree of implementation measures.

In defining the treatment measures, we need to identify the variables that are critical to the LRDC instructional model. There are seven domains of classroom characteristics which seem to be most relevant to the instructional model and in which classrooms may differ:

1. Testing procedures
2. Prescription practices
3. Traveling skills of teachers (how the teacher moves about the classroom, reinforcing appropriate student behavior)
4. Instructional materials actually used
5. Allocation of time
6. Space and its utilization
7. Teachers' knowledge of the curriculum and the children in her charge.

In moving from these domains to possibly measureable dimensions, at least two avenues are possible. In the testing procedures domain, for example, one could define ways in which the teachers might vary in testing practices, such as:

1. Frequency of individual testing
2. Accuracy of scoring and recording test results
3. Location of testing area in the classroom
4. Use of mastery level
5. Testing for all terminal objectives.

A member of the LRDC Follow-Through staff (Champagne, 1971) has defined such a list, consisting of 108 items for seven model components to be checked by a classroom observer. Tryout this past Spring indicated the promise of this procedure as a device for assessing the effectiveness of the Follow-Through teacher training program. But it appears necessary to identify a few key variables in each area if data collection and analysis is to be a manageable task for purposes of evaluation. More than 150 classrooms could be available for evaluation, and observer costs must be kept down.

Reynolds (1971) has suggested a good example of how this can be done. His studies of a few classrooms at the Oakleaf School have shown that the closer the placement and testing procedures conform to the instructional model, the higher the correlation between the child's location in the curriculum and standardized achievement scores. One key assumption of our instructional model is that learning will proceed most efficiently if the child works at that place in the curriculum which lies just above what he has mastered and below what he had not mastered. Frequent criterion-referenced testing is the mechanism whereby this placement or location is continuously determined. If this is done carelessly, the child will be wasting his time at tasks he has already mastered, or at tasks he hasn't the prerequisites to master.

For a given classroom, correlations between curriculum location or placement and general level of achievement will be low if:

1. Children are allowed or encouraged to move through the curriculum without mastering each unit.

2. Children work in the curriculum below their level of mastery.

3. Teachers curtail placement distributions by keeping the class more or less together at the same general location in the curriculum.

Thus, within-classroom correlations between Fall standardized achievement tests and the child's Fall placement in the curriculum would be a good estimate of how well the teachers handle testing in the program. The other six domains are being similarly assessed to see how a few key variables might be used to represent each degree of implementation domain.

Now that the general nature of this third set of variables has been indicated, let us return to the problem of defining an analytical scheme for studying process measures in combination with input and output. There are a number of possible approaches to this problem, four of which are:

1. Canonical correlation of input and output, then relating residuals on output factors to process measures.

2. Multiple correlation of input with each output, computing residuals on each output measure and relating that to process measures.

3. Multiple partial input from output, and then canonical between output residuals and process, yielding a multiple part canonical correlation.

4. Multiple partial input from output and process, and canonical between residuals of output and process.

One main consideration is whether variance associated with input should be partialed from both output and process or just output. It is reasonable to expect that input measures will affect process. That is, treatments might vary as a function of the location and shape of the classroom distributions on input. It certainly would be useful to know the nature of such relationships, but we are primarily interested in how the treatments actually used explain variance in output not related to input.

In order to make a first pass at that question, a multiple correlation between the four Fall input measures and the Spring means was run (Table 5), followed by the computation of residuals for Spring means, thus yielding variation in class output means not explained by the four input measures. Because of the dominance of Spring means in defining the Spring canonical factor of Table 4, the Fall multiple correlation structure is identical to the Fall canonical correlation structure, reaffirming the earlier point about the lack of additional information in the Spring distribution measures. Figure 2 shows the relation between predicted and observed means for the 57 classrooms. The residuals are the vertical distances of each classroom from the center regression line.

In order to get some suggestions from staff members who know these classrooms regarding critical dimensions of classroom differences, I created

Table 5

Predicting Spring Classroom Arithmetic Means from Fall Statistics
(N = 57 Classrooms)

Fall Quantification Predictor	Criterion Correlation	Standardized Partial Regression Coefficients	Structure
Mean	.59	.64	.82
St. dev.	.39	-.19	.54
Skewness	-.49	.63	-.68
Kurtosis	-.20	.46	-.28
	Multiple correlation = .72		

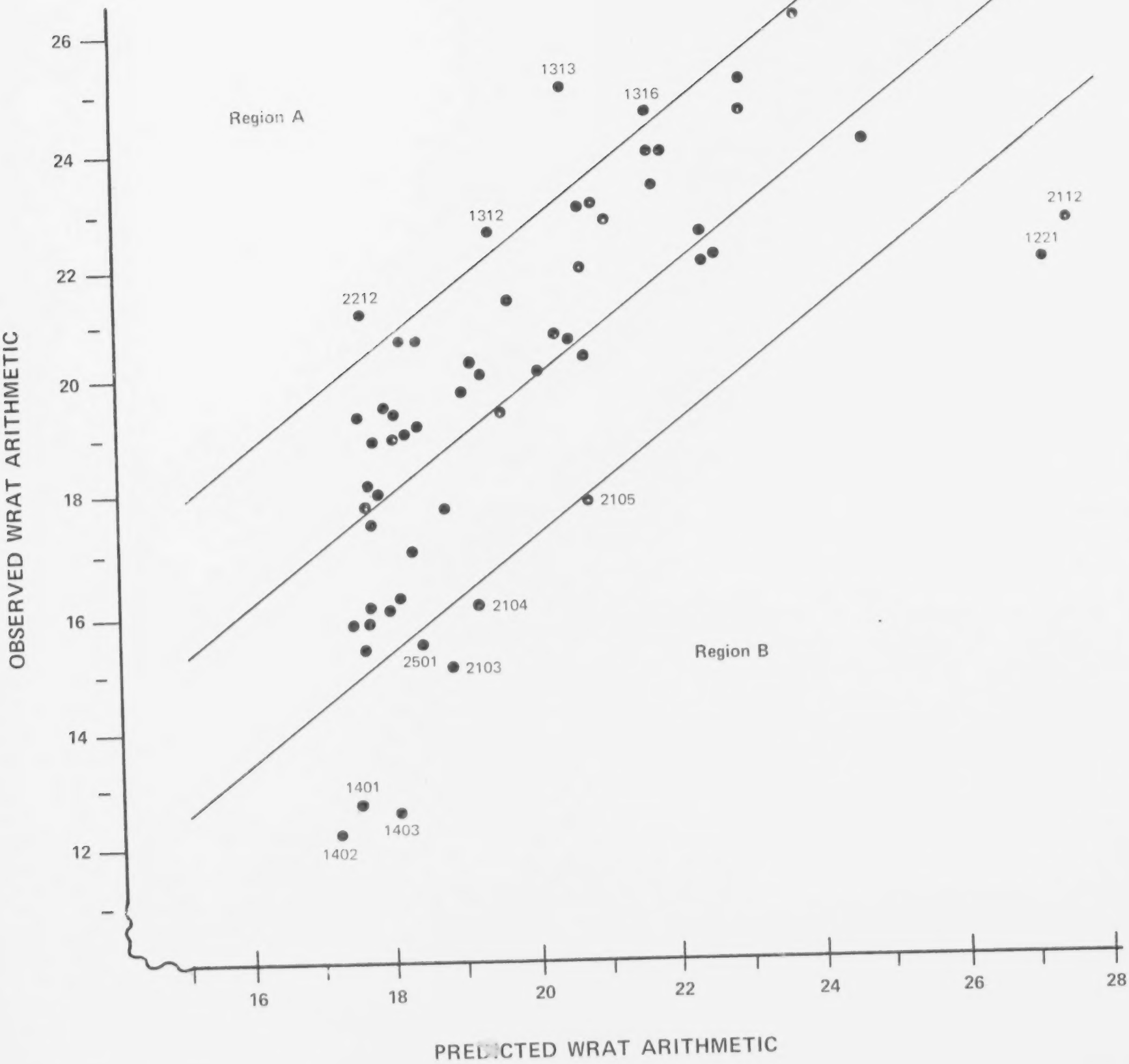

Figure 2

Location of 57 Classrooms in a 2-Space Defined by a Linear Function of Four Fall Quantification Measures (Input) and Spring WRAT Arithmetic (Output)

two lists, one consisting of classrooms with high positive residuals (region A of Figure 2), the other of classrooms with high negative residuals (region B). The two lists were not identified as such. They initially had trouble identifying differences, because classrooms in which the teacher seemed to have performed well in the traveling role and other "key" dimensions of the instructional model were on both lists, and so were less effective classrooms. However, one disturbing consistency began to appear. In region A, the teachers tended to stop placement testing prematurely, thus under-estimating the general level of entering behavior of their classrooms. Those in region B had tended to use (whenever available) the student's location in the quantification curriculum the previous Spring as the Fall placement, thus over-estimating their students because Summer retention was not taken into account.

Thus, this first pass at implementing this evaluation approach told me more about the ways in which placement testing differed among classrooms than about relations between treatments and effects. Of course, placement testing is part of the instructional model and under teacher control, but once implementation differences in that aspect of the model were detected, nothing further about the instructional model could be learned from those particular data using that regression approach.

When a researcher discovers that one of his key measure is like a rubber band, he goes back to the drawing board. Fortunately, about this same time a friend appeared on the scene with a better way to draw.

Lohnes (1971) helped me to see the need for a theory of the input and output measures which would make this research process less cut and try. This is especially important when it takes one school year for each try. Actually, Lohnes has not only shown the need for a theory of the data, he has provided us with a good one. To tell this story I need to go back a few years.

As some of you know, Lohnes and I have had the pleasure of mining the Project TALENT data, a national longitudinal study which began with over 400,000 ninth- through twelfth-graders in 1960 (Flanagan et al., 1962). A two-day battery of tests and questionnaires was administered at that time, with follow-up data collected at key points following high-school graduation. In that research, we became quite impressed with the predictive potency of a small set of orthogonal factors which Lohnes (1966) derived from the large battery of TALENT predictors. Eleven factors of abilities and motives seemed to preserve all the information available for predicting the post-high-school adjustments that we studied (Cooley & Lohnes, 1968).

When I joined LRDC, I was disappointed with their disregard for these basic, general dimensions of individual differences. Glaser (1968) and others finally convinced me that such general "aptitudes" or "motives" have little or no relevance for instructional decision making. The basic dimensions of TALENT, which are so potent as predictors of success and satisfaction in our society, are useless in determing the appropriate lessons for a child on a day to day basis.

Then last month Lohnes came along and forced me to re-examine those TALENT dimensions, not as predictors in the instructional model, but as criteria of the model. He argued that if an instructional model is worthwhile, it should increase the likelihood of the child's success and satisfaction in adult life. But as we are actively defining and redefining the model, we cannot conduct 20 year longitudinal studies between successive approximations to see how we are doing. One solution is to use these TALENT factors, these variables intervening between pre-high-school education and post-high school adjustment, as the criteria for the effectiveness of our instructional model. Of course, the TALENT battery itself is not appropriate for grade school children, but the primary factors which emerged from that battery could be found in other batteries.

Thus, this approach makes the selection of the test battery for evaluation far less arbitrary. It provides more credibility to the evaluation results because the factors have been shown to have transfer value in adult life. It also shows how it might be possible to relate elementary school practices to the process of career development, a concern recently expressed by a number of officials at the Office of Education.

One argument which has been taking place in evaluation circles (at least in ours) is the question of whether the criterion battery for evaluation should consist of standardized tests or homemade tests, limited to items which sample objectives actually in the curriculum to be evaluated. The answer seems clearer to me now.

Our own tests are important because they are needed to answer the question of whether our instructional program actually teaches the behaviors it is designed to teach. But a comprehensive evaluation effort needs to do more than that. It needs to demonstrate how well children from that program are equipped to cope after they leave that school. If primary factors of abilities and motives are good predictors of success and satisfaction as young adults, if they have face validity for the criteria they are predicting, and if those factors can be estimated by a mixture of standardized tests and measures derived from operating our instructional model, then those factors can and should be the criteria of our program's effectiveness.

A complete description of the TALENT factors requires an entire monograph (Lohnes, 1966), but it is possible to at least summarize those key factors which had predictive potency in the follow-up studies (Cooley & Lohnes, 1968). Four core ability factors emerged from 60 TALENT surface traits: verbal knowledges, English language, mathematics, and visual reasoning. The best predictor of follow-up criteria and the most important explanatory construct of the intercorrelations among the 60 TALENT ability traits is the verbal knowledges factor. Lohnes (1966) admits that this is a close approximation to general intelligence. He chose to call it verbal knowledges because "intelligence is a term that is much more susceptible to misunderstanding than is knowledge." However, the time may be ripe for us to begin to get people to realize that

one outcome of schools can and should be to maximize student scores on a general intelligence factor.[3]

From the 38 typical performance measures (interests and needs), Lohnes derived 11 motive factors, four of which were potent predictors of what people tended to do after leaving high school. Three of these factors were very familiar interest dimensions: business, science, and outdoor. The fourth motive factor was called scholasticism. Lohnes (1966) defines scholasticism as "a source motive that explains a pattern of school-sited behaviors the society approves and rewards [pp. 5-19]."

Our evaluative research this school year will be guided by the results of the past year's evaluation efforts, Lohnes' theory for the input and output measures, and the need to further define the degree of implementation domain. Next Fall we should know a little more about our instructional model than we know this Fall. Evaluative research can and must proceed as an integral part of curriculum development. It is not a one-shot exercise which one goes through when a new program is "finished." It is not easily partitionable into formative and summative activities. It can provide

---

[3]As I completed this paper I was delighted to discover that my colleagues Glaser and Resnick (1972) just finished their draft of a review of instructional psychology for the 1972 Annual Review, in which they include a discussion of the research which "treats aptitudes as the dependent rather than control variables, and seeks to influence them through instructional intervention." It should be an exciting year for us at LRDC as we try to bridge this gap between what makes good psychometric sense and what we know about instructional psychology!

information to developers while it is seeking information for potential consumers. It is research. It is facilitated by organizing ideas. It does proceed through a series of successive approximations. It is sometimes "cut and try" but it is never cut and dry.

References

Champagne, D. W. Assessment of the LRDC Follow-Through: Curriculum components and role performance. Pittsburgh: Learning Research and Development Center, 1971, in preparation.

Cooley, W. W., & Lohnes, P. R. Predicting development of young adults. Palo Alto: American Institutes for Research, 1968.

Cooley, W. W., & Lohnes, P. R. Multivariate data analysis. New York: Wiley, 1971.

Flanagan, J. C. et al. Design for a study of American youth. Boston: Houghton Mifflin, 1962.

Glaser, R. Adapting the elementary school curriculum to individual performance. Proceedings of the 1967 invitational conference on testing problems. Princeton, N.J.: Educational Testing Service, 1968. Pp. 3-36.

Glaser, R., & Resnick, L. B. Instructional psychology. Annual Review of Psychology, 1972, in preparation.

Jastak, J. F., Bijou, S. W., & Jastak, S. R. Wide Range Achievement Test. Wilmington, Del.: Guidance Association, 1965.

Lindvall, C. M., & Bolvin, J. O. Programed instruction in the schools: An application of programing principles in individually prescribed instruction. Sixty-Sixth Yearbook of the NSSE, Part II. Chicago: NSSE, 1967. Pp. 217-254.

Lindvall, C. M., & Cox, R. C. Evaluation as a tool in curriculum development: The IPI evaluation program. Chicago: Rand McNally, 1970.

Lohnes, P. R. Measuring adolescent personality. Pittsburgh: American Institutes for Research, 1966.

Lohnes, P. R. Statistical descriptors of school classes. Submitted to Journal of Educational Measurement, 1971, 15 pp.

Lohnes, P. R. Planning for evaluation of the LRDC instructional model. Pittsburgh: Learning Research and Development Center, 1971, in preparation.

Research for Better Schools. Progress report II: Individually prescribed instruction. Philadelphia: Research for Better Schools, 1971.

Resnick, L. B., Wang, M. C., & Kaplan, J. Behavior analysis in curriculum design: A hierarchically sequenced introductory mathematics curriculum. Pittsburgh: Learning Research and Development Center, 1970.

Reynolds, L. J. A strategy for the evaluation of individualization. Pittsburgh: Learning Research and Development Center, 1971, in preparation.

Wang, M. C., Resnick, L. B., & Schuetz, P. R. PEP in the Frick elementary school: Interim evaluation report 1968-1969. Pittsburgh: Learning Research and Development Center, 1970.

Wittrock, M. C., & Wiley, D. E. (Eds.) The evaluation of instruction: Issues and problems. New York: Holt, Rinehart, and Winston, 1970.